Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 4770–4774, July 1988
Evolution

Sources and evolution of human *Alu* repeated sequences

(interspersed repeat/retroposition/insertion/CpG dinucleotide)

ROY J. BRITTEN*†‡, WILL F. BARON*, DAVID B. STOUT*, AND ERIC H. DAVIDSON*

*Division of Biology, California Institute of Technology, Pasadena, CA 91125; and †Carnegie Institution of Washington, Washington, DC 20005

Contributed by Roy J. Britten, March 10, 1988

ABSTRACT ***Alu* repeated sequences arising in DNA of the human lineage during about the last 30 million years are closely similar to a modern consensus. *Alu* repeats arising at earlier times share correlated blocks of differences from the current consensus at diagnostic positions in the sequence. Using these 26 positions, we can recognize four subfamilies and the older ones are each successively closer to the 7SL sequence. It appears that there has existed a series of conserved genes that are the primary sources of the *Alu* repeat family, presumably through retroposition. These genes have probably replaced each other in overlapping relays during the evolution of primates.**

Nearly a million copies of the *Alu* repeated sequence (1) are interspersed throughout human DNA with an average spacing of ≈4 kilobases (2). *Alu* repeated sequences appear to frequently induce rearrangements, as indicated by the following example. Five different hereditary defects in the low density lipoprotein receptor gene, causing familial hypercholesterolemia, all result from deletions or duplications in which *Alu* repeated sequences occur at the rearrangement break points (3). The rearrangements occur frequently at specific regions within the *Alu* repeated sequences in the low density lipoprotein receptor gene and the γ-, δ-, and β-globin genes (3). The large number of *Alu* repeats inserted in gene regions by retroposition (4) and the events of rearrangement they cause have been major sources of variation during primate evolution.

Recently, *Alu* repeats have been shown to be divided into at least three subsets (5), including sets with a "conserved consensus" and a "divergent consensus." This paper and the accompanying paper (6) examine similar and divergent sets of *Alu* sequences by using more powerful methods. The analysis adds to previous work by identifying the diagnostic substitutions that are shared among subfamilies; by comparing the divergence of many pairs of sequences with their divergence from the consensus; by identifying correlated sets of mutations of the progenitor 7SL (7) sequence; and by hybridization measurements of the total set of *Alu* repeated sequences. We examine a model in which conserved "source" sequences are repeatedly copied and the copies are inserted into the genome.

METHODS

Alignments were done with a Wilbur and Lipman (8) program and manually checked to reach the best alignment to the consensus of Fig. 1, assuming equal weight for insertions, deletions, and mismatches.§ Repurified commercial human placental DNA was sonicated to ≈500-nucleotide fragments. To suppress the effect of base composition DNA was bound to hydroxyapatite at 50°C and thermally eluted in PT [2.0 M tetraethylammonium chloride/0.013 M neutral phosphate buffer (PB)]. Precise duplexes of this size have a t_e (half point for elution) of 68°C. A clone of a gorilla DNA region (bases 264–1043) (9) containing a recent class IV *Alu* repeat in M13 was labeled by extension and the single-stranded fragment was purified (Gor-e) or labeled by fill in of duplex replicative form (Gor-f). Because of its high G+C content, the single-stranded DNA bound to hydroxyapatite under standard conditions (0.12 M PB, 45°C) and was completely eluted only at 69°C.

```
X         10         20         30         40
GGCcgGGcgcgGTGGCTCAcgCCTGTAATCCCAGCACTTTGGGAGGCcg
                     A
50     :  60: +   ::70!   +   80    +   90:!    !:
AGGcgGGcgGATCAcgAGGTCAGGAGATcgAGACCATCCTGGCTAACAcg
                                                 A
100 ;      110        120      !  130        140
GTGAAACCCcgTCTCTACTAAAAA-TACAAAAAATTAGCcgGGcgTGGTG
                      A
   !          +        170        180        190:  !
GcgGGcgCCTGTAGTCCCAGCTACTcgGGAGGCTGAGGCAGGAGAATGGc
      -                 G
!200:    :  210      !:220          :        240
gTGAACCcgGGAGGcgGAGCTTGCAGTGAGCcgAGATcgcgCCACTGCAC
           -
 250        260        270 :        X
TCCAGCCTGGGcgACAGAGcgAGACTCcgTCTC
```

FIG. 1. Consensus of the closely similar set of *Alu* sequences. Below the line are shown the differences between the consensus and an *Alu* repeat that was inserted into the gorilla genome (9) after the split between the lineages leading to human and gorilla. The 3′-terminal poly(A) region is not shown. The 25 CpG dinucleotides are in lowercase letters. Symbols over the sequence show locations of the diagnostic positions used to distinguish classes of *Alu* repeats, as described at the end of this paper. Class: IV/III, !; III/II, +; II/I, :.

RESULTS

The primary analysis in this paper was done with 30 *Alu* repeated sequences chosen to include *Alu* sequences that are very similar to each other and to include a few moderately divergent sequences, which are not so divergent that uncertainties in alignment arise. Fig. 1 shows the consensus of these sequences, which is almost identical to the "conserved" consensus (5) mentioned above. A consensus is defined by the majority of nucleotides at each position, so the minority of divergent *Alu* repeats has no effect. Thus, the consensus of Fig. 1 is strictly for the most closely related known *Alu* repeated sequences. For brevity, it is called "the consensus" rather than "the closely related subset consensus." The fact that a number of sequences differ from the consensus only by independent random mutations indicates that these sequences have been copied from a common unchanging sequence, since if the source sequence mutated

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: SD, shared mutations (substitutions, insertions, or deletions) differing from a reference sequence.

‡To whom reprint requests should be addressed at: California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625.

§Fortran programs are available as well as complete alignments and references to be copied onto your floppy disk.

during the copying period all of the subsequent copies would share the mutation.

Shared Differences from the Consensus (SD). If a set of DNA sequences derive from a common origin, are duplicated, and are subject to changes, we may group together those with nucleotides similar to each other but differing from a progenitor sequence or a consensus sequence. Substitutions, insertions, or deletions shared among a subset but differing from the reference sequence may be informative. However, most mutations of *Alu* repeats occur repeatedly at CpG dinucleotide hot spots (10) and are not informative. Nevertheless, a minority of mutations do have strong implications regarding origin and relationship among the *Alu* repeated sequences. We need a term inclusive of both informative and noninformative positions and define SD (shared differences) as the shared mutations (substitutions, insertions, or deletions) differing from a reference sequence. SD may be used for any pair or group of sequences, for a particular position, subset of positions, or all positions. The meaning of SD differs from that of "shared derived characters" since the reference sequence used here is the modern consensus.

The SD of 30 *Alu* Sequences. The SD that occur at the same position in a large fraction of 30 *Alu* sequences are exhibited in Fig. 2. This diagram focuses on a few highly shared

```
CLASS:        IV          III     II

  93 T.............. .CCCC CCCACCCCC 12  |Diagnostic
  98 C..T.A.......T... .TTTT TTTTTTTTT 15  |
 218 C.......-...-.... GGGGG GGGGGGGGG 14  |
 196 G..........-..... CCCCC CTCCCTCCC 12  |for
 199 G..........-..... TTTTT TTTTTTTTC 13  |
 132 A.-...-........ ----- ----.---- 15  |
 152 G.............T. CC.CC CCC.AA.CT  8  |class III

64+1 -.............. ..... CTTCTCCC.  5  |Diagnostic
64+2 -.............. ..... TTTT.TTTC  7  |
  76 A.............. ..... TTTTTTTTT  9  |for
  86 T........A..... ..... GGGGGAGGG  8  |
 162 G.............. ..C.. .AAA.AAAA  7  |class II

  72 G............A... .AAAA .....A...  6   Reversion G/A/G
 153 G...-.........A.. .A.AA .AA.T.AA.  8 These 2 were CpGs
 197 G..........-..... T.A.A A.....AC.  4 in class II & III
---------------------- ----- ----------------------------
   4 C.........T..TT. .GT.. .T...T...  6
   5 G.....AA...AAA... ...A. ....A.AAA 10
  57 C....TTT......A.. T.... T.TTT..T.  9  Below line is
  58 G...........A...A .A.A. .....AA.A  7  a sample of
 138 C.........T.T... ..TTT .....TTA  7  positions with
 139 G.....AAA......AA ..... AA.AAA..A 11  large SD
 150 C..T.T.T...T....G ....T ...TA..T.  7  among the
 151 G.....A......AAA. A..A. AA..CAA.. 10  50 that
 174 C.T.T.........T... T.T.. .T.TTT..T 10  are part
 175 G.....C.......-.. ..... T.A.-.A..  2  of CpGs.
 206 C...........-T..T TT.T. .T...TTA.  8
 207 G..T...A.A.A-.... ..AA. ....A...A  7
 213 C.........T-...T ....T ....T.TTT  7
 214 G...........-AAA. C.A.A .AAA.A...  9
 236 C..T.TG..T...G..T T...T .T......T  8
 237 G...C..A..A...A.T ..AAT CA.A-AT.A  9
 238 C........T..T...- ....T .TA.TTGTT  8
 239 G.....A...A...A.. ..... A..A.A...  6
 268 C.........-T.TTA. T.... T..T....T  7
 269 G.....A.A.-.A..C. .A... ..AA.AA..  8
 276 C..........T..TT. ..-.. TT...TTT.  8
 277 G.......A.A...... AA.A. ..A.A.T.A  8
     abcdefghijklmnopq rstuv ABCDEFGHI
```

FIG. 2. Nucleotides in the diagnostic positions of 30 *Alu* repeats. The consensus nucleotide is shown at the left and the *Alu* repeats are aligned approximately in order of divergence from the consensus. All matches to consensus are shown as periods. The sequence has been rearranged so that the diagnostic positions with the largest number of SD are near the top. The numbers to the right are the number of SD from the consensus. Below the line are included a sample of CpG dinucleotide positions to show the lack of pattern in the sharing of mutations at these hot spots. Letters at bottom identify the individual *Alu* repeats. Numbers on left are the positions in the sequence of Fig. 1. The following is a list of the divergence of the *Alu* repeats from the consensus in the whole sequence, excluding the poly(A) tail, counting all insertions and deletions as 1 regardless of size: a, consensus; b, 6; c, 9; d, 10; e–g, 14; h and i, 17; j and k, 21; l, 23; m, 24; n, 26; o, 29; p, 33; q, 36; r, 28; s, 33; t, 37; u, 38; v, 39; A, 38; B, 40; C, 42; D, 43; E, 44; F, 46; G and H, 47; I, 52.

nucleotides so the sequence has been rearranged first to show off the diagnostic positions and second to isolate the CpG dinucleotides that are hot spots for mutation. There are 50 of these positions and only a sample of them has been listed below the line in Fig. 2. The *Alu* repeats that were inserted earliest (and have had a greater chance to be mutated since insertion) are shown at the right and the most recently inserted ones are shown at the left. This diagram immediately identifies diagnostic positions that define several classes of *Alu* repeats, recognized from the rows of identical nucleotides, which differ from the consensus and are shared among the more divergent (older) *Alu* repeats.

The 12 positions at the top of Fig. 2 appear to show true shared derived mutations occurring as two sets, one of five mutations and a second one including seven mutations. In each case, almost all of the nucleotides are shared among almost all of the subset of *Alu* repeats. There are 14 *Alu* repeats that share the upper seven and nine more that share these as well as five additional diagnostic nucleotides. In other words, the presence of any one of these nucleotides most often implies that the others of the set will be present, and they are effective class diagnostics. Class IV is the set of *Alu* repeats that matches the consensus at most of the 12 positions. Class III is the set that matches the consensus at the lower 5 positions but matches the older *Alu* repeats at the upper 7 positions. Class II is the set that matches the older *Alu* repeats at most of the 12 positions. Class I is discussed below and is class J in ref. 6.

Sequential Replacement of *Alu* Sources. The patterns of sharing shown at the top of Fig. 2 could in principle arise either (*i*) from a series of different sources producing each class of *Alu* repeats in turn and being sequentially replaced during evolution, or (*ii*) from a set of coexistent sources producing all classes at all times. From the start it has appeared that (*i*) is correct since the more divergent *Alu* repeats to the right on this diagram were probably inserted into the genome at earlier times. The following evidence supports this view.

Several different concurrently operating sources would produce sets of *Alu* repeats of all classes, each initially identical to the respective source. Then each class would age and would include sequences with a range of divergence from each other and from their source. However, observation shows the opposite. Members of the classes that are more divergent from the consensus are more divergent from each other. For example, the average divergence of all class II members from each other is 56.3 mutations (20% substitutions, insertions, and deletions including CpG positions for the rightmost 9 Alu repeats in Fig. 2). The average divergence from each other of the leftmost 13 *Alu* repeats of class IV is 31.6 mutations (11%). Thus, the rightmost 9 *Alu* repeats were copied from earlier source sequences and inserted into the genome far in the past and since that time have diverged from the source sequence that existed when they were inserted as well as from each other. Fig. 3 (Triangles) shows the result of similar calculations using a set of 54 sequences. We conclude that in the past the source sequences differed from the modern consensus in the two successive patterns shown at the top of Fig. 2. The more recent event changed seven positions in the source and the earlier changed five positions.

Mutations Less Shared Among *Alu* Repeats. For any small number of random independent mutations, the divergence between two sequences will be the sum of their divergences from their source sequence. Therefore, their divergence from each other is expected to be just twice the average of their divergences from the consensus, and the expected slope of the line on Fig. 3 is 1/2, while the best straight line has a slope of 0.52. Note that for these nondiagnostic non-CpG positions, the sequence of the ancient source is the same as the modern consensus. We conclude that nondiagnostic and non-CpG

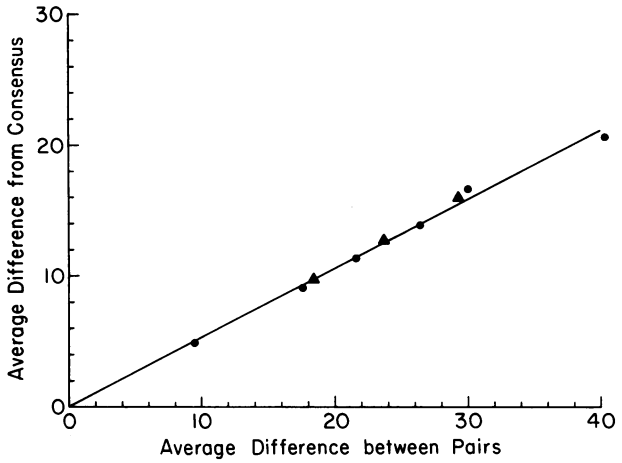

FIG. 3. Demonstration that mutations at a majority of positions occur independently. Sequences of similar divergence from the consensus (see Fig. 1) are clustered and the average divergence from the consensus is plotted against the average divergence of each from all the others in a cluster (●). Axes show the number of positions that differ in a total of 217 positions that are neither CpGs nor diagnostic positions. ▲, Clusters of class II, III, and IV in a set of 54 *Alu* sequences. The divergences of class II are underestimated since only 9 relatively young class II *Alu* sequences are in this set. The line has slope 0.52, compared to an expected value of 0.50. When the CpGs or the diagnostic positions are included, the points form a sharply rising curve because of the lack of independence of mutations at these positions.

differences from the consensus supply a good estimate of age, as does the sequence divergence among a subset of *Alu* repeats.

Rates of Change at CpGs. The lack of pattern in the sharing at most CpG positions is due primarily to the large number of repeated independent transitions at these hot spots for mutation, which are a result of methylation of the cytosines (11). Fig. 4 shows the high CpG rate. Initially, ≈3/4 of the mutations occur at the 25 CpG dinucleotides and ≈1/4 occur at the 217 non-CpG, non-diagnostic positions for an average *Alu* repeat. However, the process effectively stops at ≈33% substitution, an average loss of ≈2/3 of the CpGs. Because of the complex kinetics, CpG positions are difficult to interpret in evolutionary calculations for *Alu* sequences.

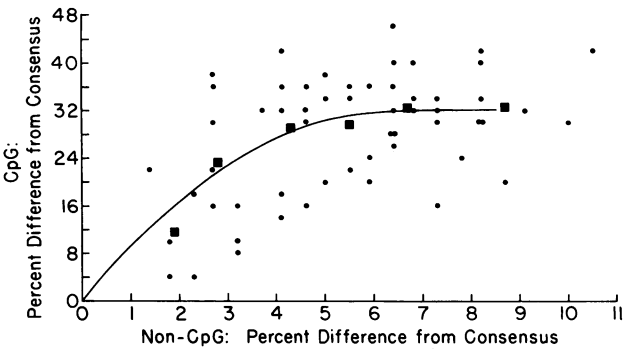

FIG. 4. Evolutionary dynamics of CpG dinucleotides after insertion of *Alu* repeated sequences. The percentage substitution of the Cs and Gs of CpG dinucleotides is plotted against the percentage substitution at 217 positions that are not CpGs and not part of the diagnostic set of positions for classes II, III, and IV (●). Data for 54 sequences including 27 class IV, 17 class III, and 9 class II. Groups of 5–10 points have been averaged to exhibit the dynamics (■). The scatter seems to imply that the mutations at CpG dinucleotides are less frequent in some genomic regions into which *Alu* sequences have been inserted. The diagram covers ≈60 million years of evolution.

Drift Rate of *Alu* Sequences After Insertion. There is evidence that the *Alu* sequence drifts at the rate of single copy DNA (12). Nine examples of *Alu* repeats in identical positions in chimpanzee and human globin gene regions are available for comparison (12–14), and the degree of divergence is 1.4% (32/9/231) based on the number of non-CpG positions in the consensus. Single copy DNA hybridization measurements (15) show a divergence of 1.6% and interspecies sequence differences of the β-globin region (16) also gives a divergence of 1.6%. *Alu* repeats in orangutan and human DNA have been compared in the β-globin gene cluster (16) and show a difference of 2.5% for non-CpG sites, while single copy DNA shows 3.4% (15). It appears the non-CpG *Alu* repeat drift rate equals the primate single copy DNA rate of ≈0.15% per million years in each lineage (15, 17).

Measurement of Most *Alu* Sequences by Hybridization. Almost all of the *Alu* repeats that have been sequenced are from gene regions and may not be a random sample of the 900,000 copies that are in the human genome. To reveal the characteristics of the majority of *Alu* repeats, a labeled probe was made from a cloned *Alu* repeat of class IV recently inserted into the gorilla genome after the human and gorilla lineages separated (9). This probe (almost identical in sequence to the consensus) was hybridized to human DNA and the thermal denaturation characteristics of the duplexes were determined. The probe Gor-e (see *Methods*) hybridizes efficiently with human DNA (data not shown) by incubation at 60°C in PT, where the melting temperature (t_m) of precise duplexes is 68°C. The product is a duplex with a t_m of 64°C, indicating that a large set of precise congener *Alu* repeats (young class IV) is present in the genome. Another probe [Prb1.8(2)], which is a member of class II, does not hybridize at all under these conditions, although it does hybridize efficiently at lower temperatures in PT, indicating that few if any precise copies of some more divergent *Alu* repeats are in the human genome.

The fraction of *Alu* sequences closely related to the consensus is estimated to be ≈25% in the measurement of

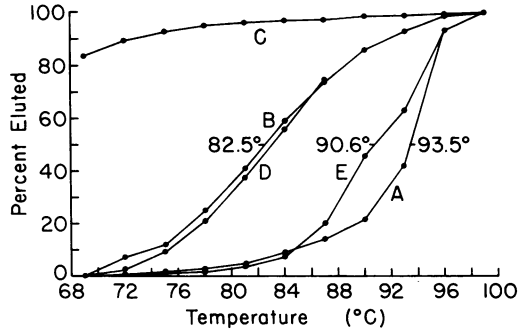

FIG. 5. Demonstration of a large set of precise class IV sequences. A 730-nucleotide labeled *Alu* probe, Gor-f (see *Methods*), was melted native (curve A). The probe was denatured and stripped of nondenaturable duplexes on hydroxyapatite and reassociated with total sheared human DNA at 50°C in 0.12 M PB to C_0t 0.5 and passed over hydroxyapatite in 0.12 M PB and eluted at 3°C intervals (curve B). To examine the high melting fraction of hybrids with human DNA, an identical run (curve D) was stopped at 87°C and the 25% of the hybrids that remained were eluted with 0.48 M PB without denaturation and were bound to hydroxyapatite a second time. The melting curve of this fraction (curve E) has a t_e (see *Methods*) of 90.6°C or 3°C below native DNA. Of the 3°C t_m reduction, 1°C is due to the reduced length of the hybrids (300 vs. 730) and 2°C could be explained by the 2% divergence of the probe from the consensus. Thus, it appears that 25% of the *Alu* sequences in the human genome are diverged from the consensus by little more than 2%. The *Alu* sequences from gene regions almost totally lack this precise fraction. For the control (curve C), the human DNA (curve B) was replaced with sea urchin DNA.

Table 1. Dominant nucleotides in the diagnostic positions

Position	Class IV		III		II		I		7SL
	88%		77%		86%		83%		
196	G	/	C		C		C	=	C
199	G	/	T		T		T	=	T
218	C	/	G		G		G	=	G
152	G	/	C		C		C	=	C
127*	A	/	—		—		—		
93	T	/	C		C		C	=	C
98	C	/	T		T		T	=	T
72	G	/	A	/	G		G	=	G
	95%		96%		80%		92%		
76	A		A	/	T		T	=	T
86	T		T	/	G		G	=	G
64 + 1	—		—	/	T		T	=	T
64 + 2	—		—	/	C	/	T	=	T
162	G		G	/	A	/	G	=	G
	89%		90%		81%		73%		
57	C		C		C	/	A	=	A
62	C		C		C	/	G	=	G
68	G		G		G	/	C	/	T
69	T		T		T	/	C	=	C
92	C		C		C	/	G	=	G
99	G		G		G	/	A	=	A
104	A		A		A	/	G	=	G
193	A		A		A	/	G	=	G
203	A		A		A	/	G	=	G
207	G		G		G	/	A	=	A
219	T		T		T	/	C	=	C
232	A		A		A	/	T	=	T
274	T		T		T	/	C	=	C
	25		17		71		31		
	(b		c		a		J)		

Representation of the sequence differences between the four known classes of *Alu* sequences, numbered in the order of their appearance during primate evolution (at the top), evolution proceeding from right to left. Listed are the dominant nucleotides in each class for each diagnostic position (see Fig. 1). Data are from ref. 6 as well as this work. Jurka and Smith (6) use the symbols in parentheses shown at the bottom to identify the classes. Slashes show the mutations that occurred as the new classes appeared. The average percentage of occurrence of the dominant nucleotide in a class is shown over the column of positions to which it applies. The number of *Alu* sequences known in each class is at the bottom. The sets of mutations between classes appear as coordinate changes at one time, although they may have been spread over short evolutionary periods. In the right column are shown the nucleotides in the human 7SL sequence; equals signs indicate that the dominant nucleotide in a class equals that in 7SL. Positions are numbered at the left as in the consensus of Fig. 1.

*Position represents a change in the length of an internal string of 5 or 6 As, not part of 7SL.

Fig. 5, for which the recent gorilla insert probe was used. The result does not agree with the known set of sequenced *Alu* repeats from gene regions. These contain only three closer than 2% divergence from the consensus (Fig. 4) or 1–2% of the *Alu* pool from which these were drawn. The implication is that gene regions do not include a fair sample of recently inserted *Alu* repeats and do not contain a random set of *Alu* sequences.

DISCUSSION

Subfamilies of *Alu* Sequences. There would be little underlying significance to subfamilies if a source sequence continuously evolved by point mutations while being copied. There would, of course, be sets of recently inserted *Alu* repeats and sets of older ones separable by their degree of divergence. The resulting broad distribution could be split into narrow fractions, each with a consensus sequence. The older groups would differ more from the modern consensus, and their members would be more divergent from each other, but the boundaries would be totally arbitrary. To discriminate against such a model, some specific coherence of subfamilies must be shown. Subsets of *Alu* repeats have been proposed, arguing from the widths of distributions of divergence (5), but without specific evidence for coherence. Nevertheless, their "divergent" subset is like class I and the "conserved" is similar to class IV, although the more divergent majority of this class could not be recognized by their method. Our evidence for subfamily structure is that particular groups of mutations are correlated. In other words, if one mutation (of a correlated set) is present in a given *Alu* sequence, there is a very high probability that the other mutations of that set will be present in that sequence. Table 1 is an assembly of the positions that define the four classes, showing the mutational events.

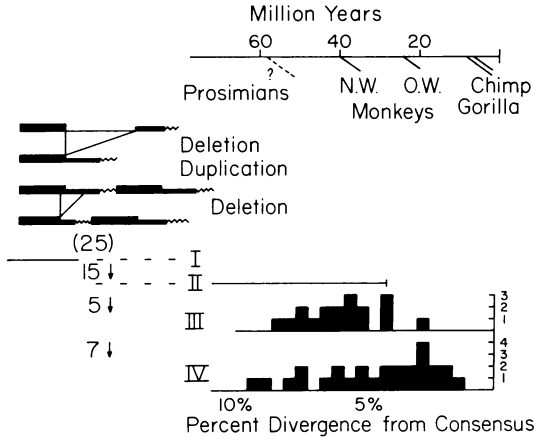

FIG. 6. Diagram of the evolution of the *Alu* sources from the progenitor 7SL sequence. At the top is an approximate time scale and the times of branching in the primate lineage are indicated for reference. At the left are shown the probable steps in the evolution of the 7SL sequence to form the class I source gene. The first step was probably deletion of ≈150 nucleotides in the center of the 7SL, then duplication, then a deletion of ≈25 nucleotides in the left half. Finally, there are ≈52 mutations required to reach the modern class IV source. Twenty-four of these are shown in Table 1, but it is not known when the other mutations occurred or whether they occurred in 7SL. The lines labeled class I and II show the approximate periods of production of these classes of *Alu* repeats, based on the number of mutations of members of these classes. Numbers beside arrows are the number of diagnostic changes between classes (Table 1). Finally, the histograms show the distribution of percentage differences from the consensus of all of the known class III and IV *Alu* repeats, for non-CpG and nondiagnostic positions. N.W., New World; O.W., Old World.

The quality of the diagnostic positions for identifying classes is indicated by the percentages listed in Table 1. Most of the 20% differences from the dominant nucleotide are random mutations and do not usually suggest misclassification. If all of the 26 diagnostic positions were tested for a new sequence, the chance for a mistaken classification is very small. For example, to discriminate between classes II and III, there are 5 positions each with a chance of 20% of not equalling the diagnostic nucleotide. Thus, the chance of failure is <0.1% in the worst case and the certainty of classification is almost always very much better. Such a high level of certainty demonstrates the reality of subfamilies.

Evolutionary History of the *Alu* Sources. The evolution of the source genes has been very different from that of the

individual *Alu* repeated sequences (after their insertion into the genome). The contrast is very clear for changes at CpG dinucleotides and for insertions or deletions. The homologous part of the 7SL gene is high in CpG (10 CpG per 148 nucleotides) and a few more have been added so that the consensus now has 25 CpGs in a length of 281 nucleotides. In contrast, Fig. 4 shows the *Alu* repeats rapidly lose CpGs. One insertion has occurred in the evolution of the sources, while among the 30 *Alu* repeats of Fig. 2, the following insertions have occurred: 11 single-nucleotide; 2 double-nucleotide; 4 four-nucleotide; 1 seven-nucleotide; and 1 eight-nucleotide insertion. The source genes have suffered 1 two-nucleotide deletion, while the set of 30 *Alu* sequences includes 67 single- or double-nucleotide deletions.

During its dominant period (Fig. 6) the source of class IV *Alu* repeats has not changed in sequence, although if it was mutated at the rate of drift observed for *Alu* repeats after insertion, a 4.5% change (10 mutations) would have occurred, exclusive of CpGs. These observations indicate that selection against sequence change has been important in the history of the source genes, suggesting that they carry out a function. A reasonable proposal would be the production of an RNA that is part of a functional ribonucleoprotein particle similar to the signal recognition particle (18), which contains 7SL RNA. If such a particle exists, the RNA would have a sequence nearly identical to the consensus. The *Alu* repeats could be thought of as nearly a million pseudogenes.

The differences of the modern human *Alu* source from the modern 7SL can be explained by the steps in Fig. 6 (*Upper Left*). It is not possible to decide about the order of many of the events, but the first deletion and creation of an *Alu*-like sequence probably occurred before the mammalian radiation, since the rodents have a similar sequence that is half the length of the modern human *Alu*. The duplication and next deletion probably occurred after the primate lineage was established and before the split with the prosimian lineage, since only primates appear to have the double-length *Alu*, which is also present in prosimians. A galago sequence (19) shares the basic structure of the human *Alu* and 7 of the defining nucleotides of class I, indicating that the transition from class I to class II occurred in the primate lineage after the split with the prosimians. It is possible to date the four classes of *Alu* repeats from the number of mutations (differing from the consensus) in the positions that are not diagnostic positions or CpGs. The average values for classes III and IV are 6.4% and 4.5%, suggesting that their sources were active 43 and 30 million years ago. Fig. 6 shows the distribution of age estimates for *Alu* repeats of classes III and IV suggests overlap in the times that the different sources were active. Class III apparently ceased activity about the time that the average class IV sequence was inserted into the genome, ≈30 million years ago. It is certain that class IV was not produced in a short burst of activity (5), since one recent member of class IV (9) was inserted into the gorilla genome after the gorilla and human lineages separated, while the oldest known members of this class were probably inserted ≈60 million years ago, as indicated in Fig. 6.

The data indicate that a series of genes has existed that have been sequentially derived from each other but nevertheless coexisted. The mutations that are diagnostic for the change from each class to the subsequent class are almost all retained right up to the present, as shown in Table 1. As each new class of source evolved it was more divergent from the progenitor 7SL sequence by virtue of sets of coordinated changes.

We thank Barbara Barth for sequence entry and manuscript aid; Jerzy Jurka for data before publication and for additional III and IV class *Alu* repeats; Carl Schmid for prepublication copies of alignments and references during the review process (5); G. Trabuchet for a clone of a gorilla *Alu* repeat; and J. L. Goldstein, M. A. Lehrman, and D. W. Russell for unpublished *Alu* sequences. This work was supported by National Institutes of Health Grant GM34031.

1. Deininger, P. L. & Schmid, C. W. (1979) *J. Mol. Biol.* **127,** 437–460.
2. Hwu, H. R., Roberts, J. W., Davidson, E. H. & Britten, R. J. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3875–3879.
3. Lehrman, M. A., Russell, D. W., Goldstein, J. L. & Brown, M. S. (1987) *J. Biol. Chem.* **262,** 3354–3361.
4. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55,** 631–661.
5. Willard, C., Nguyen, H. T. & Schmid, C. W. (1987) *J. Mol. Evol.* **26,** 180–186.
6. Jurka, J. & Smith, T. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 4775–4778.
7. Ullu, E. & Tschudi, C. (1984) *Nature (London)* **312,** 171–172.
8. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
9. Trabuchet, G., Chebloune, Y., Savatier, P., Lachuer, J., Faure, C., Verdier, G. & Nigon, V. M. (1987) *J. Mol. Evol.* **25,** 288–291.
10. Bains, W. (1986) *J. Mol. Evol.* **23,** 189–199.
11. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. (1978) *Nature (London)* **274,** 775–780.
12. Sawada, I., Willard, C., Shen, C.-K. J., Chapman, B., Wilson, A. C., & Schmid, C. W. (1985) *J. Mol. Evol.* **22,** 316–322.
13. Maeda, N., Bliska, J. B. & Smithies, O. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5012–5016.
14. Sibley, C. G. & Ahlquist, J. E. (1984) *J. Mol. Evol.* **20,** 2–15.
15. Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987) *Science* **238,** 369–373.
16. Koop, B. F., Goodman, M., Xu, P., Chan, K. & Slightom, J. L. (1986) *Nature (London)* **319,** 234–238.
17. Britten, R. J. (1986) *Science* **231,** 1393–1398.
18. Blobel, W. (1982) *Nature (London)* **299,** 691–698.
19. Daniels, G. R. & Deininger, P. L. (1983) *Nucleic Acids Res.* **11,** 7595–7610.